APPENDIX A

SENSITIVITY ANALYSIS: HOW DO WE KNOW WHAT'S IMPORTANT?

A.0 INTRODUCTION

Sensitivity analysis, as it is applied to risk assessment, is any systematic, common sense technique used to understand how risk estimates and, in particular, risk-based decisions, are dependent on variability and uncertainty in the factors contributing to risk. In short, sensitivity analysis identifies what is "driving" the risk estimates. It is used in both point estimate and probabilistic approaches to identify and rank important sources of variability as well as important sources of uncertainty. The quantitative information provided by sensitivity analysis is important for guiding the complexity of the analysis and communicating important results (see Chapter 6). As such, sensitivity analysis plays a central role in the tiered process for PRA (see Chapter 2). This Appendix focuses on a set of graphical and statistical techniques that can be used to determine which variables in the risk model contribute most to the variation in estimates of risk. This variation in risk could represent variability, uncertainty, or both, depending on the type of risk model and characterization of input variables.

There is a wide array of analytical methods that may be referred to as sensitivity analysis, some of which are very simple and intuitive. For example, a risk assessor may have two comparable studies from which to estimate a reasonable maximum exposure (RME) for childhood soil ingestion. One approach to evaluating this uncertainty would be to calculate the corresponding RME risk twice, each time using a different plausible point estimate for soil ingestion rate. Similarly, in a probabilistic model, there may be uncertainty regarding the choice of a probability distribution. For example, lognormal and gamma distributions may be equally plausible for characterizing variability in an input variable. A simple exploratory approach would be to run separate Monte Carlo simulations with each distribution in order to determine the effect that this particular source of uncertainty may have on risk estimates within the RME range (90^{th} to 99.9^{th} percentile, see Chapter 1).

Sensitivity analysis can also involve more complex mathematical and statistical techniques such as correlation and regression analysis to determine which factors in a risk model contribute most to the variance in the risk estimate. The complexity generally stems from the fact that multiple sources of variability and uncertainty are influencing a risk estimate at the same time, and sources may not act independently. An input variable contributes significantly to the output risk distribution if it is both highly variable *and* the variability propagates through the algebraic risk equation to the model output (i.e., risk). Changes to the distribution of a variable with a high sensitivity could have a profound impact on the risk estimate, whereas even large changes to the distribution of a low sensitivity variable may have a minimal impact on the final result. Information from sensitivity analysis can be important when trying to determine where to focus additional resources. The choice of technique(s) should be determined by the information needs for risk management decision making.

This appendix presents guidance on both practical decision making and theoretical concepts associated with the sensitivity analysis that are commonly applied in risk assessment. An overview of the type of information provided by sensitivity analysis is presented first, followed by guidance on how to decide what method to use in each of the tiers. A straightforward example of applications of Tier 1 and Tier 2 sensitivity analysis methods is shown, followed by a more detailed discussion of the theory and equations associated with the different methods.

EXHIBIT A-1

DEFINITIONS FOR APPENDIX A

Continuous Variables - A random variable that can assume any value within an interval of real numbers (e.g., body weight).

Correlation - A quantitative expression of the statistical association between two variables; usually represented by the Pearson correlation coefficient for linear models, and the Spearman rank correlation coefficient (see below) for nonlinear models.

Discrete Variables - A random variable that can assume any value within a finite set of values (e.g., number of visits to a site in one year) or at most a countably infinite set of values, meaning that you can count observations, but there is no defined upper limit. An example of countably infinite would be the number of dust particles in a volume of air (a Poisson distribution), whereas *uncountably* infinite would be the number of points in a line segment.

Local Sensitivity Analysis - Evaluation of the model sensitivity at some nominal points within the range of values of input variable(s).

Monte Carlo Analysis (MCA) or Monte Carlo Simulation - The process of repeatedly sampling from probability distributions to derive a distribution of outcomes. MCA is one of several techniques that may be used in PRA.

Multiple Regression Analysis - A statistical method that describes the extent, direction, and strength of the relationship between several (usually continuous) independent variables (e.g., exposure duration, ingestion rate) and a single continuous dependent variable (e.g., risk).

Nonparametric Tests - Statistical tests that do not require assumptions about the form of the population probability distribution.

Range Sensitivity Analysis - Evaluation of the model sensitivity across the entire range of values of the input variable(s).

Rank - If a set of values is sorted in ascending order (smallest to largest), the rank corresponds to the relative position of a number in the sequence. For example, the set {7, 5, 9, 12} when sorted gives the following sequence {5, 7, 9, 12} with ranks ranging from 1 to 4 (i.e., rank of 5 is 1, rank of 7 is 2, rank of 9 is 3, and rank of 12 is 4).

Sensitivity Analysis - Sensitivity generally refers to the variation in output of a model with respect to changes in the values of the model's input(s). Sensitivity analysis attempts to provide a ranking of the model inputs based on their relative contributions to model output variability and uncertainty. Common metrics of sensitivity include:

- Pearson Correlation Coefficient - A statistic r that measures the strength and direction of linear association between the values of two quantitative variables. The square of the coefficient (r^2) is the fraction of the variance of one variable that is explained by least-squares regression on the other variable, also called the coefficient of determination..
- Sensitivity Ratio - Ratio of the change in model output per unit change in an input variable; also called *elasticity*.
- Sensitivity Score - A sensitivity ratio that is weighted by some characteristic of the input variable (e.g., variance, coefficient of variation, range).
- Spearman Rank Order Correlation Coefficient - A "distribution free" or nonparametric statistic r that measures the strength and direction of association between the ranks of the values (not the values themselves) of two quantitative variables. See Pearson (above) for r^2.

A.1.0 UTILITY OF SENSITIVITY ANALYSIS

As highlighted in Exhibit A-2, sensitivity analysis can provide valuable information for both risk assessors and risk management decision makers throughout the tiered process for PRA. By highlighting important sources of variability and uncertainty in the risk assessment, sensitivity analysis is generally an important component of the overall uncertainty analysis. For example, methods that quantify parameter uncertainty and model uncertainty may yield different estimates of the RME risk. This information can be used to guide the tiered process by supporting decisions to conduct additional analyses or prioritize resource allocations for additional data collection efforts. Results of sensitivity analysis can also facilitate the risk communication process by focusing discussions on the important features of the risk assessment (e.g., constraints of available data, state of knowledge, significant scientific issues, and significant policy choices that were made when alternative interpretations of data existed).

EXHIBIT A-2

UTILITY OF SENSITIVITY ANALYSIS

- **Decision making with the tiered approach**- *e.g., After quantifying parameter uncertainty, we are 95 percent confident that the RME risk is below the risk level of concern— no further analysis is needed.* Also—*selection of a beta distribution over a lognormal distribution for ingestion rate changes the 95th percentile of the risk distribution by a factor of 10—further evaluation may be needed.*
- **Resource allocation** - *e.g., Two of the 10 exposure variables contribute 90 percent of the variability in the risk estimate.*
- **Risk communication** - *e.g., For input variable X, if we were to use a distribution based on site-specific data instead of a national survey, we would expect a minimal change in the RME risk estimate.*

Decision Making with the Tiered Approach

In general, the type of information provided by a sensitivity analysis will vary with each tier of a PRA. Table A-1 provides an overview of the methods that may be applied in each tier based on the type of information needed. In Tier 1, sensitivity analysis typically involves changing one or more input variables or assumptions and evaluating the corresponding changes in the risk estimates. Ideally, the results for Tier 1 would be useful in deciding which exposure pathways, variables, and assumptions are carried forward for further consideration in subsequent tiers of analysis. By identifying the variables that are most important in determining risk, one can also decide whether point estimates, rather than probability distribution functions (PDFs), can be used with little consequence to the model output. This information is important not only for designing 1-D MCA models of variability, but also for designing more complex analyses of uncertainty discussed in Appendix D (e.g., 2-D MCA models, geostatistical analysis, Bayesian analysis). Section A.2.2 provides an overview of the Tier 1 methods and some insights regarding their limitations. Methods associated with Monte Carlo simulations used in Tiers 2 and 3 can take advantage of the ability to vary multiple inputs simultaneously and account for correlations. Sections A.2.3 and A.3 provide an overview of the sensitivity analysis methods that can be applied in a probabilistic analysis.

Table A-1. Overview of Sensitivity Analysis Methods Applicable in Tiers 1, 2, and 3 of a PRA.

Tier	Goal	SA Method(s)	What to Look For	Rationale
1	Quantify contributions of each exposure pathway to risk, identify major and minor pathways	Calculate % of total risk from each exposure pathway	Exposure pathways that contribute a very small percentage (e.g., < 5%) to total risk	Good preliminary step in Tier 1 for reducing the number of exposure variables to focus on in subsequent tiers.
			Exposure variables that appear in multiple exposure pathways	Risk estimates are likely to be more sensitive to variables that appear in multiple exposure pathways.
1	Identify the form of the dose equation for key pathways	Inspection	Equation is multiplicative or additive	SR values can be determined with minimal effort (see Table A-3). For multiplicative equations, SR=1.0 for all variables in the numerator, and SR is a function of the percent change for all variables in the denominator.
			Equation contains variables with exponents (e.g., powers, square roots)	Output is likely to be more sensitive to variables with exponents greater than 1.0.
1	Quantify contributions of each exposure variable to total risk, identify major and minor variables	Sensitivity Ratio (SR), unweighted	SR = 1.0, or SR is the same for multiple variables	It's likely that this is a multiplicative equation (see above), and the SR approach will not be effective at discriminating among relative contributions. Explore sensitivity further with other methods.
			SR ≠ 1.0	SR may vary as a function of the % change in the input variable. In this situation, it can be informative to explore small deviations (± 5%) and large deviations (min, max) in the input variables.
			SR < 1.0	Implies an inverse relationship between the input and output variables (e.g., inputs in the denominator of a risk equation).

Table A-1. Overview of Sensitivity Analysis Methods Applicable in Tiers 1, 2, and 3 of a PRA.

Tier	Goal	SA Method(s)	What to Look For	Rationale
			SR=0	Variable probably appears in both the numerator and denominator and, therefore, cancels out of the risk equation. Examples include exposure duration (ED) in noncancer risk equations, and body weight (BW) if ingestion rate is expressed as a function of body weight.
1	(cont'd) Quantify contributions of each exposure variable to total risk	Sensitivity Ratio (SR), weighted—also called Sensitivity Score	Differences in SR based on the weighting factor	A more informative approach than unweighted SR value for those variables that have sufficient information to define a weighting factor (e.g., coefficient of variation or range).
2	Quantify relative contributions of exposure pathways to risk	1-D MCA for variability or uncertainty, with outputs specifying % contribution of exposure pathways	Compare mean with high- and low-end percentiles of % contribution to risk	The % contribution of each exposure pathway will vary as a function of the variability (or uncertainty) in the inputs; exposure pathways that appear to be relatively minor contributors on average, or from Tier 1 assessment, may in fact be a major contributor to risk under certain exposure scenarios. The likelihood that a pathway is nonnegligible (e.g., > 5%) can be useful information for risk managers.
2	Quantify relative contributions of exposure variables to risk	1-D MCA for variability or uncertainty, Graphical analysis—scatterplots of inputs and output	Nonlinear relationship	Easy and intuitive approach that may identify relationships that other methods could miss. May suggest transformations of input or output variables (e.g., logarithms, power transformations) that would improve correlation and regression analyses.
		1-D MCA, Correlation Analysis using Pearson and /or Spearman Rank	Very high or low correlation coefficients Differences between relative rankings based on Pearson and Spearman	Easy to implement with commercial software; rank orders the variables based on the *average* contribution to variance. Differences in magnitude of coefficients are expected between Pearson and Spearman rank approaches, but relative order of importance is likely to be the same.

Table A-1. Overview of Sensitivity Analysis Methods Applicable in Tiers 1, 2, and 3 of a PRA.

Tier	Goal	SA Method(s)	What to Look For	Rationale
		1-D MCA, Multiple Linear Regression Analysis (e.g., stepwise)	Very high or low regression coefficients R^2 and adjusted R^2 for total model	Easy to implement with commercial software; gives contribution to reduction in residual sum of squares (RSS) For risk equations with large sets of input variables, a small subset of inputs may be able to explain the majority of the variance.
2	Quantify relative contributions of exposure variables to RME risk range	1-D MCA; same as previous step, but for subset of risk distribution (e.g., > 90th percentile)	Difference in relative contributions for entire risk distribution and the RME range of the risk distribution	Variables may contribute differently to the high-end of the risk distribution, especially if the input variables are highly skewed. This situation would warrant a closer look at the assumptions regarding the estimate of the variance, differences in the upper tail (high-end percentiles) for alternative choices of probability distributions, and assumptions associated with truncation limits.
		1-D MCA, Goodness-of-fit, K-S or Chi-square; Sort output as above; perform GoF on input distribution only, comparing subset of input values corresponding with high-end risk to subset corresponding with remainder of risk distribution	GoF result—rejection of null (distributions are the same) suggests the variable may be an important contributing factor to the RME risk estimate	A second method for identifying variables that contribute differently at the high-end of the risk distribution. GoF test results should be interpreted with caution because a Monte Carlo simulation will generally yield large sample sizes (e.g., n=5,000 iterations), which is more likely to result in a positive GoF test (i.e., rejection of the null).
3	Quantify relative contributions of exposure pathways and variables to variability and uncertainty in risk	2-D MCA, same sensitivity analysis methods as Tier 2	For variability, evaluate inner loop values; for parameter uncertainty, evaluate outer loop values	The results of a sensitivity analysis depend on the question that is being asked about the risk estimate—are we interested in variability or uncertainty? The major sources of variability in risk may point to a different set of input variables than the major sources of uncertainty in risk.

Resource Allocation

Decisions regarding allocation of future resources and data collection efforts to reduce lack of knowledge generally should take into consideration the most influential input factors in the model, and the cost of gaining new information about the factors. Sensitivity analysis is a key feature of determining the expected value of information (EVOI) (see Appendix D). Once a sensitivity analysis is used to identify an input variable as being important, the source of its variability generally should be determined. If an input factor has a significant uncertainty component, further research and/or data collection can be conducted to reduce this uncertainty. Reducing major sources of uncertainty, such as the most relevant probability model for variability or the parameter estimates for the model, will generally improve confidence in the model output, such as the estimated 95th percentile of the risk distribution. An input factor may contribute little to the variability in risk, but greatly to the uncertainty in risk (e.g., the concentration term). Likewise, a variable may contribute greatly to the variability in risk, but, because the data are from a well characterized population, the uncertainty is relatively low (e.g., adult tap water ingestion rate).

An example of the output from a 2-D MCA of uncertainty and variability (see Appendix D) is shown in Figure A-1. Assume for this example that the decision makers choose the 95th percentile risk as the RME risk, and that a sensitivity analysis is run to identify and quantitatively rank the important source(s) of parameter uncertainty. The bar chart (top panel) in Figure A-1 indicates that the mean soil concentration contributes most to the uncertainty in the 95th percentile risk estimate. In addition, the mean exposure frequency is a greater source of uncertainty than the standard deviation exposure frequency. Since both the sample size and variance impact the magnitude of the confidence limits for an arithmetic mean soil concentration, one way to reduce the confidence limits (i.e., the uncertainty) would be to collect additional soil samples. As shown by the box-and-whisker plots (bottom panel) in Figure A-1, increasing the sample size (from n=25 to n=50) reduced the 90% confidence limits for the 95th percentile risk to below 1E-05, assuming the additional observations support the same estimate of the mean and standard deviation as the original sample.

Although the uncertainty in a risk estimate can be reduced by further data collection if the sensitive input distribution represents uncertainty, this is not necessarily true for input distributions that represent variability. For example, variability in the distribution of body weights can be better characterized with additional data, but the coefficient of variation (i.e., standard deviation divided by the mean) will not in general be reduced.

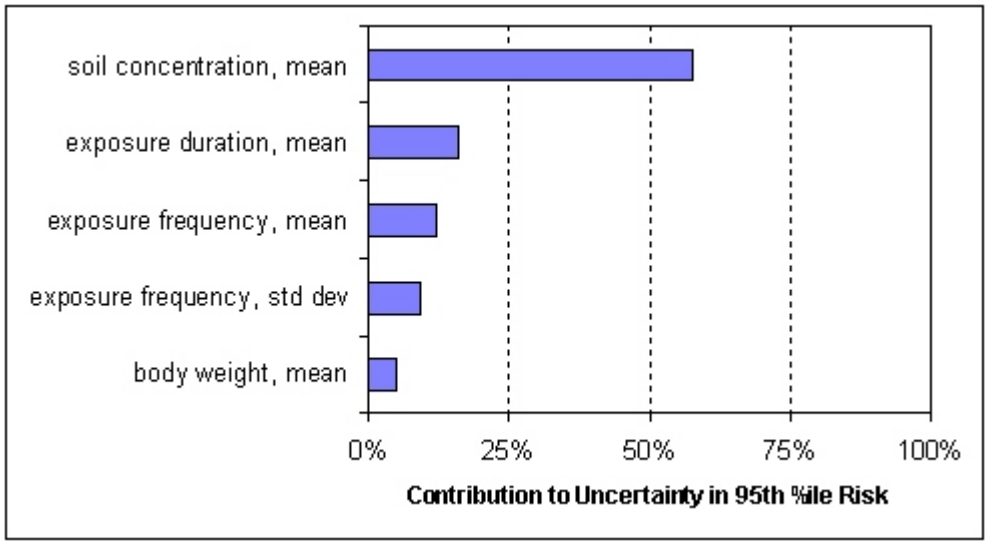
Risk Communication

Even if additional data are not collected to reduce uncertainty, identifying the exposure factors that contribute most to risk or hazard may be useful for risk communication. For example, assume that the input for exposure frequency has the strongest effect on the risk estimate for a future recreational open space. Further examination of this exposure variable reveals that the wide spread (i.e., variance) of the PDF is a result of multiple users (e.g., mountain bikers, hikers, individuals who bring picnics, etc.) of the open space who may spend very different amounts of time recreating. As a result of this analysis, the decision makers and community may decide to focus remediation efforts on protecting the high-risk subpopulation that is expected to spend the most time in the open space.

After determining which contaminants, media, and exposure pathways to carry into a PRA, numerical experiments generally should be performed to determine the sensitivity of the output to various

distributions and parameter estimates that may be supported by the available information. Variables that do not strongly affect the risk estimates may be characterized with point estimates without significantly altering the risk estimates. This guidance document does not recommend a quantitative metric or rule of thumb for determining when a variable strongly affects the output; this would generally be determined on a case-by-case basis. A qualitative or quantitative analysis may be used depending on the complexity of the risk assessment at this point. For example, incidental ingestion of soil by children is often an influential factor in determining risk from soil, a factor recognized by risk assessors. This recognition is a *de facto* informal sensitivity analysis. An array of quantitative techniques is also available, ranging from something as simple as comparing the range of possible values (i.e., maximum-minimum) for each variable, to more complex statistical methods such as multiple regression analysis. Several of these methods are discussed in more detail in this appendix.

Often, sufficient information is available to characterize a PDF for a minor variable without significant effort. This situation raises a question of whether the variable should be characterized with a point estimate or a PDF. The results of sensitivity analysis should be viewed as supplemental information, rather than an absolute rule for determining when to use a PDF. There are at least two issues to consider related to risk communication. First, the risk communication process may be facilitated by narrowing the focus of the evaluation to the key factors. More attention can be given to the discussion of key variables quantified by PDFs by describing the minor variables with point estimates. However, the decision to use a point estimate should be balanced by considering a second issue regarding perception and trust. There may be a concern that by reducing sources of variability to point estimates, there would be a reduction (however small) in the variability in risk, especially if multiple small sources of variability add up to a nonnegligible contribution. To address these concerns, it may be prudent to leave the PDFs in the calculations despite the results of a sensitivity analysis.



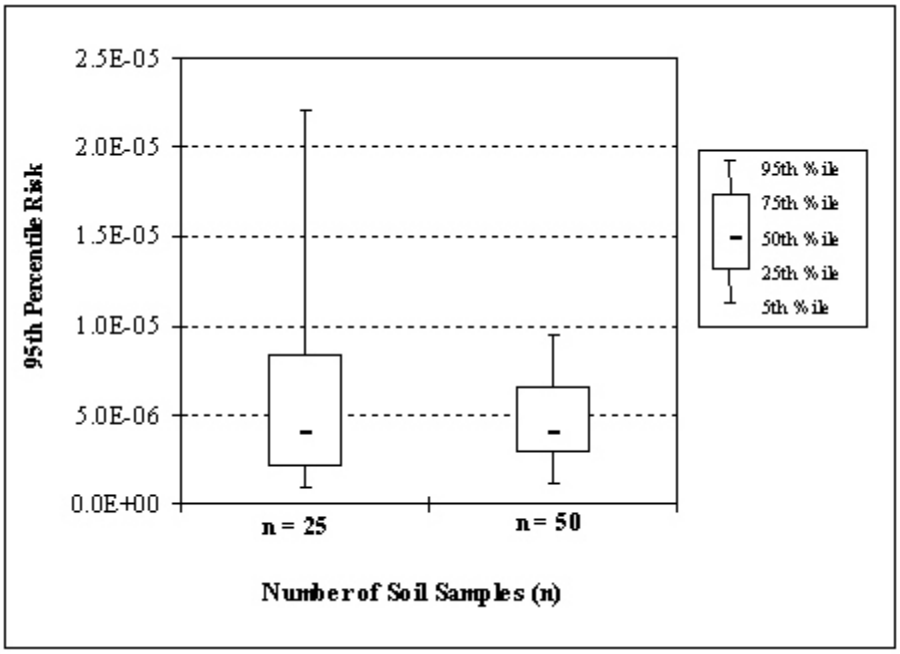


Figure A-1. Results of 2-D MCA in which parameters of input distributions describing variability are assumed to be random values. Results of a sensitivity analysis (top graph) suggest that more than 50% of the uncertainty in the 95th percentile of the risk distribution is due to uncertainty in the arithmetic mean concentration in soil. The bottom graph gives box-and-whisker plots for the 95th percentile of the risk distribution associated with Monte Carlo simulations using different sample sizes (n=25 and n=50). For this example, the whiskers represent the 5th and 95th percentiles of the distribution for uncertainty, otherwise described as the 90% confidence interval (CI). For n=25, the 90% CI is [1.0E-06, 2.2E-05]; for n=50, the 90% CI is reduced to [1.2E-06, 9.5E-06]. While increasing n did not change the 50th percentile of the uncertainty distribution, it did provide greater confidence that the 95th percentile risk is below $1x10^{-5}$.

A.2.0 COMMON METHODS OF SENSITIVITY ANALYSIS

Of the numerous approaches to sensitivity analysis that are available (see Exhibit A-3), no single approach will serve as the best analysis for all modeling efforts. Often, it will make sense to apply multiple approaches. The best choice(s) for a particular situation will depend on a number of factors, including the nature and complexity of the model and the resources available. A brief description of the more common approaches is provided in this appendix. Sensitivity analysis need not be limited to the methods discussed in this guidance, which focuses on the more common approaches. A large body of scientific literature on various other methods is available (e.g., Iman et al., 1988, 1991; Morgan and Henrion, 1990; Saltelli and Marivort, 1990; Rose et al., 1991; Merz, Small, and Fischbeck, 1992; Shevenell and Hoffman 1993; Hamby, 1994; U.S. EPA, 1997). Any method used, however, generally should be documented clearly and concisely. This includes providing all information needed by a third party to repeat the procedure and corroborate the results. Relevant information might include the following: exposure pathways and equations; a table with the input variables with point estimates, probability distributions and parameters; and tables or graphs giving the results of the sensitivity analysis and description of the method used. A hypothetical example is presented in this appendix to illustrate how to apply and present the results of selected approaches to sensitivity analysis.

EXHIBIT A-3

SOME KEY INDICES OF SENSITIVITY ANALYSIS

- Relative contribution of exposure pathways
- Inspection of risk equation
- Sensitivity ratios (i.e., elasticity)
- Sensitivity scores (i.e., weighted sensitivity ratios)
- Graphical techniques with results of Monte Carlo simulations (e.g., scatter plots)
- Correlation coefficient (or coefficient of determination, r^2) (e.g., Pearson product moment, Spearman rank)
- Normalized multiple regression coefficient
- Goodness-of-fit test for subsets of the risk distribution

Hypothetical Example of a Noncancer Risk Equation

To illustrate the application of sensitivity analysis concepts to Tier 1 and Tier 2, a hypothetical risk assessment is presented based on the general equation for Hazard Index (HI) given by Equation A-1. Note that HI is equal to the sum of the chemical-specific Hazard Quotient (HQ) values, so technically, this example reflects exposures from a single chemical.

$$HI = \frac{C_i \times I_i \times AF_i \times EF \times ED}{BW \times AT} \times \frac{1}{RfD} \qquad \text{Equation A-1}$$

The terms in Equation A-1 can be defined as follows: concentration in the i^{th} exposure medium (C_i), ingestion or inhalation rate of the i^{th} exposure medium (I_i), absorption fraction of chemical in the i^{th} exposure medium (AF_i), exposure duration (ED), exposure frequency (EF), body weight (BW), averaging time (AT=ED x 365 days/year), and reference dose (RfD).

For this example, HI is calculated as the sum of the exposures to adults from two exposure pathways: tap water ingestion and soil ingestion. Equation A-2 gives the equation for HI while Table A-2 gives the inputs for a point estimate assessment and a probabilistic assessment of variability.

$$HI = \frac{((C_w \times I_w \times AF_w) + (C_s \times I_s \times AF_s)) \times EF \times ED}{BW \times AT} \times \frac{1}{RfD} \quad \text{Equation A-2}$$

Table A-2. Point estimates and probability distributions for input variables used in the hypothetical example of HI associated with occupational exposure via water and soil ingestion.

Input Variable in Equation A-2	Point Estimate		Probability Distribution		Units
	CTE	RME	Type	Parameters	
Concentration in Water (C_w)	40	40	point estimate	40	mg/L
Tap Water Ingestion Rate (I_w)	1.3	2.0	lognormal[1]	[1.3, 0.75]	L/day
Absorption Fraction Water (AF_w)	0.30	0.50	beta[2]	[2.0, 3.0]	unitless
Concentration in Soil (C_s)	90	90	point estimate	90	mg/kg
Soil Ingestion Rate (I_s)	0.05	0.10	uniform	[0, 0.13]	kg/day
Absorption Fraction Soil (AF_s)	0.10	0.30	beta[2]	[1.22 , 4.89]	unitless
Exposure Frequency (EF)	250	350	triangular	[180, 250, 350]	days/yr
Exposure Duration (ED)	1	7	empirical[3]	see below	years
Body Weight (BW)	75	75	lognormal[1]	[74.6, 12.2]	kg
Averaging Time (AT)	365	2555	empirical[4]	ED x 365	days
RfD_{oral}[5]	0.5	0.5	point estimate	0.5	mg/kg-day

[1]Parameters of lognormal distribution are [arithmetic mean, standard deviation].
[2]Parameters of beta distribution are [alpha, beta], with range defined by min=0 and max=1.0. Parameter conversions for arithmetic mean and standard deviation are given in Table A-7.
[3]Parameters of empirical cumulative distribution function (ECDF) for ED ~ [min, max, {x}, {p}] = [0, 30, {0.08, 0.18, 0.30, 0.44, 0.61, 0.84, 1.17, 1.72, 3.1, 6.77, 14.15, 23.94}, {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.975, 0.99}], where *x* is the array of values and *p* is the array of corresponding cumulative probabilities.
[4]AT=ED x 365 for noncarcinogenic risks (Hazard Index).
[5]For simplicity, RfD_{oral} is assumed to be applicable to the ingestion of the chemical in both water and soil.

A.2.1 TIER 1 APPROACHES

Approaches for sensitivity analysis in Tier 1 of a PRA are limited to calculations that are based on changing point estimates. They are generally easy to perform and to communicate. As given by Table A-1, goals for the sensitivity analysis in Tier 1 include quantifying the relative contributions of the exposure pathways, identifying potential nonlinear relationships that may exist between input variables and the risk estimate, and rank ordering the relative contribution of exposure variables to variability or

uncertainty in the risk estimate. This last goal may be the most difficult to achieve due to the limitations associated with the point estimate methodology. Methods are applied to the hypothetical example presented above (Section A.2.0) in order to demonstrate the inherent limitations of the Tier 1 approaches in some situations.

A.2.1.1 Percentage Contribution of Exposure Pathways to Total Risk

For cancer and noncancer risk assessments central tendency exposure (CTE) and RME risk is typically calculated as the sum of risks from multiple exposure pathways. Risks may be dominated by one or two exposure pathways, which can be determined through a simple calculation as shown below. The relative contributions of exposure pathways are likely to differ between the CTE risk and RME risk.

The point estimates in Table A-2 were applied to Equation A-2 to obtain CTE and RME point estimates of HI. Table A-3 gives the percent contributions of soil ingestion and tap water ingestion using Equations A-3 and A-4. Tap water ingestion contributes at least 90% to HI, and the total HI is greater than 1.0 for both CTE and RME point estimates. If 1.0 is the level of concern for HI, and a decision was made to explore variability and uncertainty in a probabilistic analysis, this result might support prioritizing the evaluation of data and assumptions associated with the tap water ingestion pathway.

Table A-3. Percent contribution of exposure pathways to HI for the example in Section A.2.

Exposure Pathway	CTE Point Estimate		RME Point Estimate	
	HI	% of total[2]	HI	% of total
Soil Ingestion	0.02	6 %	0.15	13 %
Tap Water Ingestion	0.28	94 %	1.02	87 %
Total	0.30	100 %	1.17	100 %

[1]Equation A-3: $HI_{total} = HI_{soil} + HI_{water}$
[2]Example using Equation A-4: % of total RME HI for soil ingestion = (0.15 / 1.17) x 100% = 13%.

$$HI_{total} = \sum_{i=1}^{n} HI_i \qquad \text{Equation A-3}$$

$$Percent\ Contribution_i = \frac{HI_i}{HI_{total}} \times 100\% \qquad \text{Equation A-4}$$

In this example, the choice of CTE and RME point estimates reflects an effort to explore variability in HI, rather than uncertainty. Even if the concentration terms represent the upper confidence limit on the mean (e.g., 95% UCL), the point estimates chosen to represent the CTE and RME for other exposure variables reflect assumptions about the variability in exposures. There is uncertainty that the choices actually represent the central tendency and reasonable maximum exposures. To explore this uncertainty, alternative choices for CTE and RME may have been selected. This type of exploration of uncertainty in Tier 1 may also be viewed as a form of sensitivity analysis. The percent contribution of exposure pathways could be recalculated, and the sensitivity ratio approaches discussed below may also be applied.

A.2.1.2 Inspection of Risk Equation

For many Superfund risk assessments, risk equations can be characterized as relatively simple algebraic expressions involving addition, multiplication, and division of input variables. The term "product-quotient" model is often applied to describe equations such as Equation A-1. For these risk equations, the input variables that are likely to contribute most to the variability or uncertainty in risk can be identified by inspection. In addition, inspection of the risk equation can help to identify which sensitivity analysis methods are unlikely to reveal the relative importance of the input variables. This concept is illustrated by comparing the results of the sensitivity ratio approach (Section A.2.1.3) with the Tier 2 approaches (Section A.2.2) applied to the hypothetical example in Section A.2.0.

Some risk equations can be more complex, involving conditional probabilities, or expressions with exponents (e.g., $y=x^2$, or $y=\exp(1-x)$). In these cases, the Tier 1 sensitivity analysis methods may be effective and highlighting the variables that contribute most to the risk estimates.

A.2.1.3 Sensitivity Ratio (SR)

A method of sensitivity analysis applied in many different models in science, engineering, and economics is the **Sensitivity Ratio** (SR), otherwise know as the *elasticity* equation. The approach is easy to understand and apply. The ratio is equal to the percentage change in output (e.g., risk) divided by the percentage change in input for a specific input variable, as shown in Equation A-5.

$$SR = \frac{\left(\frac{Y_2 - Y_1}{Y_1}\right) \times 100\%}{\left(\frac{X_2 - X_1}{X_1}\right) \times 100\%} \qquad \text{Equation A-5}$$

where, Y_1 = the baseline value of the output variable using baseline values of input variables
Y_2 = the value of the output variable after changing the value of one input variable
X_1 = the baseline point estimate for an input variable
X_2 = the value of the input variable after changing X_1

Risk estimates are considered most sensitive to input variables that yield the highest absolute value for SR. The basis for this equation can be understood by examining the fundamental concepts associated with partial derivatives (see Section A.3.2). In fact, SR is equivalent to the normalized partial derivative (see Equation A-12).

Sensitivity ratios can generally be grouped into two categories—local SR and range SR. For the local SR method, an input variable is varied by a small amount, usually ±5% of the nominal (default) point estimate, and the corresponding change in the model output is observed. For the range sensitivity ratio method, an input variable is varied across the entire range (plausible minimum and maximum values). Usually, the results of local and range SR calculations are the same. When the results differ, the risk assessor can conclude that different exposure variables are driving risk near the high-end (i.e., extreme tails of the risk distribution) than at the central tendency region.

Demonstration of the Limitations of SR Approach

Although SR is a relatively simple and intuitive approach, it does not provide useful information under certain conditions for the more common risk equations. To demonstrate the limitations, first Equation A-5 is applied to the hypothetical example given in Section A.2.0. The results are then extended to a more general case of any of the more common risk models that involve the products of terms (i.e., multiplicative model) or the sum of terms (i.e., additive model).

Table A-4 presents an example of the local SR and range SR approach applied to the set of RME inputs given in Table A-2. For the local SR, each input was increased by 5% (i.e., Δ=+5%), while for the range SR, each input was increased by 50%. Inputs for exposure frequency were truncated at the maximum value of 365 days/year, which represents a 4.29% increase over the nominal RME value of 350 days/year.

Table A-4. Results of the Sensitivity Ratio (SR) approach applied to the hypothetical example of RME HI given in Section A.2.0. Includes *both* soil ingestion and tap water ingestion pathways.

Input Variable , X in Equation A-2[1]	Nominal RME value (X_1)	Local SR (Δ = + 5.0%)			Range SR (Δ = + 50% or max)		
		X_2	Δ in HI (%)	SR	X_2	Δ in HI (%)	SR
Tap Water Ingestion Rate, I_w (L/day)	2.0	2.1	4.35	0.87	3.0	43.5	0.87
Absorption Fraction Water, AF_w (unitless)	0.50	0.525	4.35	0.87	0.75	43.5	0.87
Soil Ingestion Rate, I_s (kg/day)	0.100	0.105	0.65	0.13	0.150	6.5	0.13
Absorption Fraction Soil, AF_s (unitless)	0.30	0.315	0.65	0.13	0.45	6.5	0.13
Exposure Frequency, EF (days/yr)	350	365[2]	4.29	1.00	365[2]	4.29	1.00
Exposure Duration, ED (years)	7	7.35	0.00	0.00	10.5	0.00	0.00
Body Weight, BW (kg)	75	78.75	- 4.46	- 0.89	112.5	- 33.33	- 0.67

[1]Only input variables that represent variability are included. Concentrations are point estimates of uncertainty. Averaging time is a function of exposure duration. RfD is a fixed point estimate.
[2]Maximum EF of 365 days/yr represents a 4.29% change in the nominal RME value of 350 days/yr.

The following observations can be made from these results:

- In decreasing order of sensitivity:

  Local SR (Δ = 5%) rankings: EF > BW > I_w = AF_w > I_s = AF_s > ED

  Range SR (Δ = 50%) rankings: EF > I_w = AF_w > BW > I_s = AF_s > ED

- EF is the most sensitive variable with an SR value of 1.0. Since EF is a variable in the numerator for both exposure pathways, this result is to be expected, as will be explained below.

- ED yields an SR=0, suggesting it does not contribute to the HI estimate. Upon closer inspection of the risk equation, it is apparent that ED occurs in the numerator of Equation A-2, as well as in the denominator (AT=ED x 365). Thus, ED effectively cancels out of the product quotient model and does not effect the estimate of HI.
- BW, the only variable in the denominator of the risk equation, is also the only variable to yield a different SR value when comparing the local and range SR approaches. Thus, BW is the only variable for which SR depends on the percent change in the input (Δ).
- BW is the only negative SR value, indicating that HI and BW are inversely related. This is true in general for any variable in the denominator of a product quotient model.
- For variables unique to the water ingestion pathway (I_w, AF_w), SR=0.87. Similarly, for variables unique to the soil ingestion pathway (I_s, AF_s), SR=0.13 These SR values are exactly the same as the percent contributions of the tap water ingestion pathway and soil ingestion pathway to HI (see Table A-3).

Since tap water ingestion is the dominant pathway (i.e., 87% of RME HI), a reasonable strategy for the Tier 1 sensitivity ratio approach might be to limit the subsequent probabilistic analysis in Tier 2 to the tap water ingestion pathway; so that input variables unique to the soil ingestion pathway would be characterized by point estimates. For this relatively simple example, this would mean that soil ingestion rate (I_s) and absorption fraction from soil (AF_s) would be described by point estimates instead of PDFs. The question to address would then become—Of the exposure variables in the tap water ingestion pathway, which ones contribute most to HI? A sensitivity ratio approach was applied to the tap water ingestion pathway to address this question. The results are presented in Table A-5.

Table A-5. Results of the Sensitivity Ratio (SR) approach applied to the hypothetical example of RME HI given in Section A.2.0. Includes *only* tap water ingestion pathway.

Input Variable , X in Equation A-2[1]	Nominal RME value (X_1)	Local SR (Δ = + 5.0%)			Range SR (Δ = + 50% or max)		
		X_2	Δ in HI (%)	SR	X_2	Δ in HI (%)	SR
Tap Water Ingestion Rate, I_w (L/day)	2.0	2.1	5.0	1.00	3.0	50	1.00
Absorption Fraction Water, AF_w (unitless)	0.50	0.525	5.0	1.00	0.75	50	1.00
Exposure Frequency, EF (days/yr)	350	365[2]	4.29	1.00	365[2]	4.29	1.00
Exposure Duration, ED (years)	7	7.35	0.00	0.00	10.5	0.00	0.00
Body Weight, BW (kg)	75	78.75	- 4.46	- 0.89	112.5	- 33.33	- 0.67

[1]Only input variables that represent variability are included. Concentrations are point estimates of uncertainty. Averaging time is a function of exposure duration. RfD is a fixed point estimate.
[2]Maximum EF of 365 days/yr represents a 4.29% change in the nominal RME value of 350 days/yr.

The following observations can be made from these results:

- In decreasing order of sensitivity:

  Local SR (Δ = 5%) rankings: I_w = AF_w = EF > BW > ED

  Range SR (Δ = 50%) rankings: I_w = AF_w = EF > BW > ED

- SR values for variables in the numerator (I_w, AF_w, and EF) are all equal to 1.0, so the SR approach suggests that they contribute equally to the HI estimate.
- BW values are the same as in Table A-4. They are negative, and the values change as a function of the percent change in the nominal RME value (Δ).

Tables A-4 and A-5 suggest that the SR approach provides essentially the same information about sensitivity as other Tier 1 methods. Specifically, inspection of the risk equation reveals that ED does not contribute to HI. In addition, for pathway-specific variables in the numerator, like ingestion rates and absorption fractions, SR values are equal to the percent contributions of the exposure pathways. This actually reflects the fact that each factor in the numerator of a multiplicative equation has an SR of 1.0.

The results of the SR approach applied to the example above can be generalized to all multiplicative and additive risk equations, as discussed below.

Generalizing the Limitations of the SR Approach

In many cases, the general equation for SR (Equation A-5) will give values that can be determined *a priori*, without doing many calculations. To understand why this is true, it is useful to simplify the algebraic expression given by Equation A-5. Let Δ equal the percentage change in the input variable, X_1. For *SR* calculations, Δ may be either positive or negative (e.g., ±5% for local SR; ±100% for range SR), and the new value for the input variable (i.e., X_2) is given by Equation A-6.

$$\begin{aligned} X_2 &= X_1 + (X_1 \times \Delta) \\ &= X_1 \times (1 + \Delta) \end{aligned} \qquad \text{Equation A-6}$$

Therefore, the denominator in Equation A-5 reduces to Δ:

$$\frac{X_2 - X_1}{X_1} = \frac{X_1(1+\Delta) - X_1}{X_1} = \frac{(1+\Delta) - 1}{1} = \Delta$$

and Equation A-5 reduces to Equation A-7:

$$SR = \frac{1}{\Delta} \times \left(\frac{Y_2 - Y_1}{Y_1} \right) \qquad \text{Equation A-7}$$

Equation A-7 can be used to evaluate SR for different types of exposure models in which the intake equation is generally expressed as a simple algebraic combination of input variables. Solutions to SR calculations for input variables in both multiplicative and additive equations are given in Table A-6. For any such risk equation, the solution will fall into one of the five categories given by Exhibit A-4.

EXHIBIT A-4

CATEGORIES OF SOLUTIONS FOR SENSITIVITY RATIOS OF MULTIPICATIVE OR ADDITIVE EQUATIONS

Case 1 SR is a constant (e.g., 1.0). SR is independent of the choice of nominal (default) values for input variables and the choice of Δ.

Case 2 SR is a constant determined only by the nominal values for the input variables. SR is independent of the choice of Δ.

Case 3 SR is constant determined only by the choice of Δ. SR is independent of the nominal values for the input variables.

Case 4 SR is a function of both the nominal values for the input variables and the choice of Δ.

Case 5 SR is 0. The variable does not contribute to the risk estimate.

Table A-6. Examples of algebraic solutions to Sensitivity Ratio calculations for additive and multiplicative forms of risk equations.[1, 2]

Equation Type (Output = Y, Inputs = A, B, C, D)		SR_A =	SR_B =	SR_C =	SR_D =
1) Additive in Numerator	$Y = \frac{A + B}{C}$	$\frac{A}{A + B}$	$\frac{B}{A + B}$	$-\frac{1}{1+\Delta}$	NA[3]
2) Additive in Denominator	$Y = \frac{A}{C + D}$	1.0	NA	$-\frac{C}{C(1+\Delta)+D}$	$-\frac{D}{D(1+\Delta)+C}$
3) Multiplicative in Numerator	$Y = \frac{A \times B}{C}$	1.0	1.0	$-\frac{1}{1+\Delta}$	NA
4) Multiplicative in Denominator	$Y = \frac{A}{C \times D}$	1.0	NA	$-\frac{1}{1+\Delta}$	$-\frac{1}{1+\Delta}$

[1]Sensitivity Ratio for input variable A for an equation that is additive in the numerator: $SR_A = A / (A + B)$.
[2]Δ=% change in input variable. For example, Δ for $C=[(C_2 - C_1)/C_1] \times 100\%$, where C_1=the original point estimate and C_2=the modified point estimate. Similarly, $C_2=C_1 (1 + \Delta)$.
[3]NA=not applicable because the variable is not in the equation.

The following observations can be made for the four forms of the risk equation, based on one of the five cases described in Exhibit A-4:

(1) Additive in Numerator

- **Case 2**: SR values for variables in the numerator depend exclusively on the nominal point estimates for all variables in the numerator. The values are independent of the choice of percent change in the inputs (Δ).
- **Case 3**: SR values for variables in the denominator depend exclusively on Δ, and are negative (i.e., inversely related to the output). Also, the lower the choice for Δ, the higher the resulting SR values. Therefore, SR is somewhat arbitrary, especially for the range SR approach since input variables may have different plausible minimum and maximum values.

(2) Additive in Denominator

- **Case 1**: SR values for variables in the numerator are always equal to 1.0. Since they are independent of the nominal values and Δ, there is no way to distinguish the relative contributions to the output.
- **Case 4**: SR values for variables in the denominator are a function of both the nominal values of variables in the denominator and Δ.

(3) Multiplicative in Numerator and (4) *Multiplicative in Denominator*

- **Case 1**: SR values for variables in the numerator are always equal to 1.0. Since they are independent of the nominal values and Δ, there is no way to distinguish the relative contributions to the output.
- **Case 3**: SR values for variables in the denominator depend exclusively on the Δ, and are negative (i.e., inversely related to the output). Also, the lower the choice for Δ, the higher the resulting SR values. Therefore, SR is somewhat arbitrary, especially for range SR since input variables may have different plausible minimum and maximum values.

These generalized results highlight a major limitation in the use of the SR approach for obtaining information from sensitivity analysis. For simple exposure models in which the relationship between exposure and risk is linear (e.g., multiplicative), the ratio offers little information regarding the relative contributions of each input variable to the risk estimate. In many cases, all of the input variables will have the same constant, either equal to 1.0 (in the case of a single exposure pathway) or equal to the relative contributions of the exposure pathways. For more complex models that combine additive, multiplicative, and nonlinear relationships between inputs and outputs (e.g., environmental fate and transport models, pharmacokinetic models), the ratio is likely to be an effective screening tool for identifying potentially influential input variables and assumptions.

Another difficulty with the SR approach is that it generally requires an assumption that the input variables are independent. Two variables may actually be positively correlated (e.g., high values of X_1 correspond with high values of X_2) or negatively correlated (e.g., high values of X_1 correspond with low values of X_2). If input variables are correlated, holding the value for one variable fixed while allowing the other to vary may produce misleading results, especially with the range sensitivity ratio approach. For example, it may not be realistic to hold body weight fixed at a central tendency while allowing skin

surface area to vary from the minimum to maximum values. An improvement over the sensitivity ratio approach would be to allow correlated input variables to vary simultaneously.

A.2.1.4 SENSITIVITY SCORE

A variation on the sensitivity ratio approach may provide more information from a Tier 1 sensitivity analysis, but it requires that additional information be available for the input variables. The *sensitivity score* is the SR weighted by a normalized measure of the variability in an input variable (U.S. EPA, 1999). Examples of normalized measures of variability include the coefficient of variation (i.e., standard deviation divided by the mean) and the normalized range (i.e., range divided by the mean), as given by Equation A-8.

$$Sensitivity\ Score = SR \times \frac{\sigma}{\mu} \quad or \quad SR \times \frac{(max - min)}{\mu} \qquad \text{Equation A-8}$$

By normalizing the measure of variability (i.e., dividing by the mean), this method effectively weights the ratios in a manner that is independent of the units of the input variable, and provides a more robust method of ranking contributions to the risk estimates than the SR alone. This approach does require that the coefficient of variation or range can be calculated for each variable. Tables A-7 and A-8 present the results of the sensitivity scores based on the CV applied to the hypothetical example from Section A.2.0.

Table A-7. Calculation of coefficient of variation (CV = SD / Mean) for the hypothetical example of RME HI given in Section A.2.0.

Input Variable , X in Equation A-2[1]	Probability Distribution[2]	Mean[3]	SD[3]	CV = SD/Mean
Tap Water Ingestion Rate, I_w (L/day)	lognormal (1.3, 0.75)	1.3	0.75	0.58
Absorption Fraction, Water, AF_w (unitless)	beta (2.0, 3.0)	0.4	0.2	0.50
Soil Ingestion Rate, I_s (kg/day)	uniform (0, 0.13)	0.065	0.038	0.58[2]
Absorption Fraction, Soil, AF_s (unitless)	beta (1.22, 4.89)	0.20	0.15	0.75
Exposure Frequency, EF (days/yr)	triangular (180, 250, 350)	260	35	0.13[3]
Exposure Duration, ED (years)	empirical CDF (see Table A-2 for parameters)	1.75	3.86	2.21
Body Weight, BW (kg)	lognormal (74.6, 12.2)	74.6	12.2	0.16

[1]Only input variables that represent variability are included. Concentrations are point estimates of uncertainty. Averaging time is a function of exposure duration. RfD is a fixed point estimate.
[2]Beta (a, b): mean=a / (a+b) and SD = ((a x b) / [(a + b)^2 x (a+b+1)])^0.5)
Uniform (min, max): mean = (min + max)/2 and SD = ((1/12)^0.5) x (max - min) = 0.289 x (max - min)
Triangular (min, mode, max): mean = (min + mode + max)/3 and SD = (1/18) x (min^2 + mode^2 + max^2 - min x max - min x mode - mode x max)
Empirical CDF ({x}, {p}): mean and SD were estimated by Monte Carlo simulation.
[3]Mean=arithmetic mean; SD=arithmetic standard deviation

Table A-8. Results of the Sensitivity Score (Score) approach applied to the hypothetical example of RME HI given in Section A.2.0. Calculations for Sensitivity Ratio (SR) and Coefficient of Variation (CV) are given in Table A-4 and Table A-7, respectively.

Input Variable , X in Equation A-2[1]	Nominal RME value (X_1)	CV (Table A-7)	Local SR (Δ = + 5%)		Range SR (Δ = + 50%)	
			SR (Table A-4)	Score[2]	SR (Table A-4)	Score[2]
Tap Water Ingestion Rate, I_w (L/day)	2.0	0.58	0.87	**0.50**	0.87	**0.50**
Absorption Fraction, Water, AF_w (unitless)	0.50	0.50	0.87	**0.44**	0.87	**0.44**
Soil Ingestion Rate, I_s (kg/day)	0.100	0.58	0.13	**0.06**	0.13	**0.06**
Absorption Fraction, Soil, AF_s (unitless)	0.30	0.75	0.13	**0.10**	0.13	**0.10**
Exposure Frequency, EF (days/yr)	350	0.13	1.00	**0.13**	1.00	**0.13**
Exposure Duration, ED (years)	7	2.21	0.00	**0**	0.00	**0**
Body Weight, BW (kg)	75	0.16	- 0.89	**- 0.14**	- 0.67	**- 0.11**

[1]Only input variables that represent variability are included. Concentrations are point estimates of uncertainty. Averaging time is a function of exposure duration. RfD is a fixed point estimate.
[2]Score=SR x CV (see Equation A-8)

The following observations can be made from these results:

- In decreasing order of sensitivity:

  - Score based on local SR (Δ = 5%): I_w > AF_w > BW > EF > AF_s > IR_s > ED

  - Score based on range SR (Δ = 50%): I_w > AF_w > EF > BW > AF_s > IR_s > ED

- Compared with the SR approach alone in which sensitivity can only be expressed for exposure pathways, the sensitivity score approach provides a measure of sensitivity for exposure variables within each exposure pathway.
- Although ED has the highest CV, it continues to have no contribution to the HI.
  - If Tier 1 sensitivity analysis is based on the sensitivity score, the highest ranked variables are generally those with the highest CV in the exposure pathway that contributes the most to the total risk (HI). For this hypothetical example, I_w and AF_w are the two highest ranked variables.

A.2.2 TIER 2 APPROACHES

Approaches for sensitivity analysis in Tier 2 of a PRA utilize the results of Monte Carlo simulations, which allows multiple input variables to vary simultaneously. The methods are relatively simple to perform with spreadsheets or commercial statistical software. The results are generally easy to communicate, although the details of the methodology are more complex than Tier 1 approaches. As given by Table A-1, goals for the sensitivity analysis in Tier 2 are the same as Tier 1:quantifying the relative contributions of the exposure pathways, identifying potential nonlinear relationships that may exist between input variables and the risk estimate, and rank ordering the relative contribution of exposure variables to variability or uncertainty in the risk estimate. In addition, since the output is a distribution, Tier 2 sensitivity analysis methods can also utilize graphical techniques to observe nonlinear relationships, as well as evaluate potential changes in relative importance of variables and assumptions for risks in the RME risk range. Methods are applied to the hypothetical example presented in Section A.2.0 in order to demonstrate the advantages over the Tier 1 methods.

A.2.2.1 GRAPHICAL TECHNIQUES

Simple scatter plots of the simulated input and output (e.g., risk vs. exposure frequency, or risk vs. arithmetic mean soil concentration) can be used to qualitatively and quantitatively evaluate influential variables. A "tight" best-fit line through the scatter plot, as indicated by the magnitude of the r^2, suggests that a variable may significantly influence the variance in risk. Hypothetical scatter plots used to identify sensitive and insensitive variables are shown in Figure A-2. Another method for visualizing the relationship between all of the inputs and outputs is to generate a scatterplot matrix (Helsel and Hirsch, 1992). This graphic shows both histograms and scatter plots for all variables on the same page.

Figure A-3 illustrates scatter plots for the 1-D MCA simulations associated with the example from Section A.2.0. Based on the r^2 values (i.e., coefficient of determination for simple linear regression analysis), the relationship between HI and I_w is very strong ($r^2 = 0.47$) while the relationship between HI and I_s is very weak ($r^2 < 0.01$), suggesting that HI is more sensitive to variability in I_w than I_s.

A.2.2.2 CORRELATION COEFFICIENTS

The variance in a risk estimate from a Monte Carlo simulation is due to the variance in the probability distributions used in the risk equation. It is commonly said that a Monte Carlo model propagates sources of variability simultaneously in a risk equation. Numerous statistical techniques, known collectively as correlation analysis and regression analysis, can be applied to a linear equation to estimate the relative change in the output of a Monte Carlo simulation based on changes in the input variables. Examples of metrics of sensitivity include the simple correlation coefficient, the rank correlation coefficient, and a variety of coefficients from multiple regression techniques. The underlying assumptions associated with these approaches are discussed in greater detail in Section A.3. As explained in Section A.3.3.1, correlation coefficients and regression coefficients are based on different interpretations of the input variables, but they can be calculated with similar equations.

When the output distribution is compared with the distribution for one input variable at a time, two of the more common approaches are to calculate the Pearson product moment correlation and the Spearman rank correlation. Correlation analysis with one input variable will generally yield reasonable results when the input variables are sampled independently in a Monte Carlo simulation. Some statistical packages offer the correlation coefficient as an index of sensitivity, so it is important to identify which

coefficient is being calculated. *Crystal Ball®* and *@Risk* can be used to calculate the Spearman rank correlation, which tends to be more robust when the relationships between inputs and outputs are nonlinear. If the relationships are linear, such as with the product quotient models presented in this appendix, the two metrics of correlation will yield similar rankings of input variables. Rank correlation coefficients shown in *Crystal Ball®* and *@Risk* are calculated by the standard method provided in most statistics texts. *Crystal Ball®* also indicates that sensitivity can be determined as contribution to variance. This is not the relative partial sum of squares techniques discussed in Section A.3.3.2 (Equation A-19). Instead, *Crystal Ball®* calculates the contribution to the variance by squaring the rank correlation coefficients and normalizing them to 100%. Many other commonly used commercial software packages will perform Spearman rank correlation. Pearson product moment correlations (r) can be calculated in Microsoft Excel using the trendline feature in a scatter plot chart, or by using the function *Correl(X array, Y array)*, where *X array* corresponds with the Monte Carlo simulation of an input variable, and *Y array* corresponds with the output of the simulation.

Figure A-4 illustrates results of the correlation analysis for the 1-D MCA simulations associated with the example from Section A.2.0. The graphics were generated using *Crystal Ball®* 2000. The results are summarized in Table A-9. If the model output variable (e.g., HI) and input variable are highly correlated, it means that the output is sensitive to that input variable. By squaring the coefficient, the results can be expressed in terms of the percentage contribution to variance in the output (Figure A-4, top panel). To determine if the correlation is positive or negative, the correlation coefficient should not be squared (Figure A-4, bottom panel). For risk equations, in general, variables in the numerator of the equation (ingestion rate, absorption fraction, exposure frequency, etc.) will tend to be positively correlated with risk, while variables in the denominator (body weight) will tend to be negatively correlated with risk. The greater the absolute value of the correlation coefficient, the stronger the relationship.

Table A-9. Results of Tier 2 sensitivity analyses applied to hypothetical example in Section A.2.0: Pearson product moment correlations and Spearman rank correlations.[1]

Exposure Variable	Product Moment Correlation		Spearman Rank Correlation[2]		
	r	r^2 x 100%	r	r^2 x 100%	normalized r^2 x 100%
Tap Water Ingestion Rate, I_w (L/day)	0.644	41.4	0.603	36.3	39.5
Absorption Fraction Water, AF_w (unitless)	0.583	34.0	0.666	44.4	48.3
Body Weight, BW (kg)	- 0.216	4.7	- 0.229	5.2	5.7
Exposure Frequency, EF (days/yr)	0.174	3.0	0.167	2.8	3.0
Absorption Fraction Soil, AF_s (unitless)	0.109	1.2	0.149	2.2	2.4
Soil Ingestion Rate, I_s (g/day)	0.061	0.4	0.099	1.0	1.1
Exposure Duration, ED (years)	0.010	0.0	0.010	0.0	0.0

[1]Monte Carlo simulation using *Crystal Ball® 2000*, Latin Hypercube sampling, and 5000 iterations.

[2]*Crystal Ball® 2000* output includes Spearman rank correlations, r, and *normalized* r^2 values, calculated by dividing each r^2 value by the sum of all the r^2 values (i.e., 0.920 in this example). Figure A-4 illustrates the r and normalized r^2 values for the Spearman rank correlation analysis.

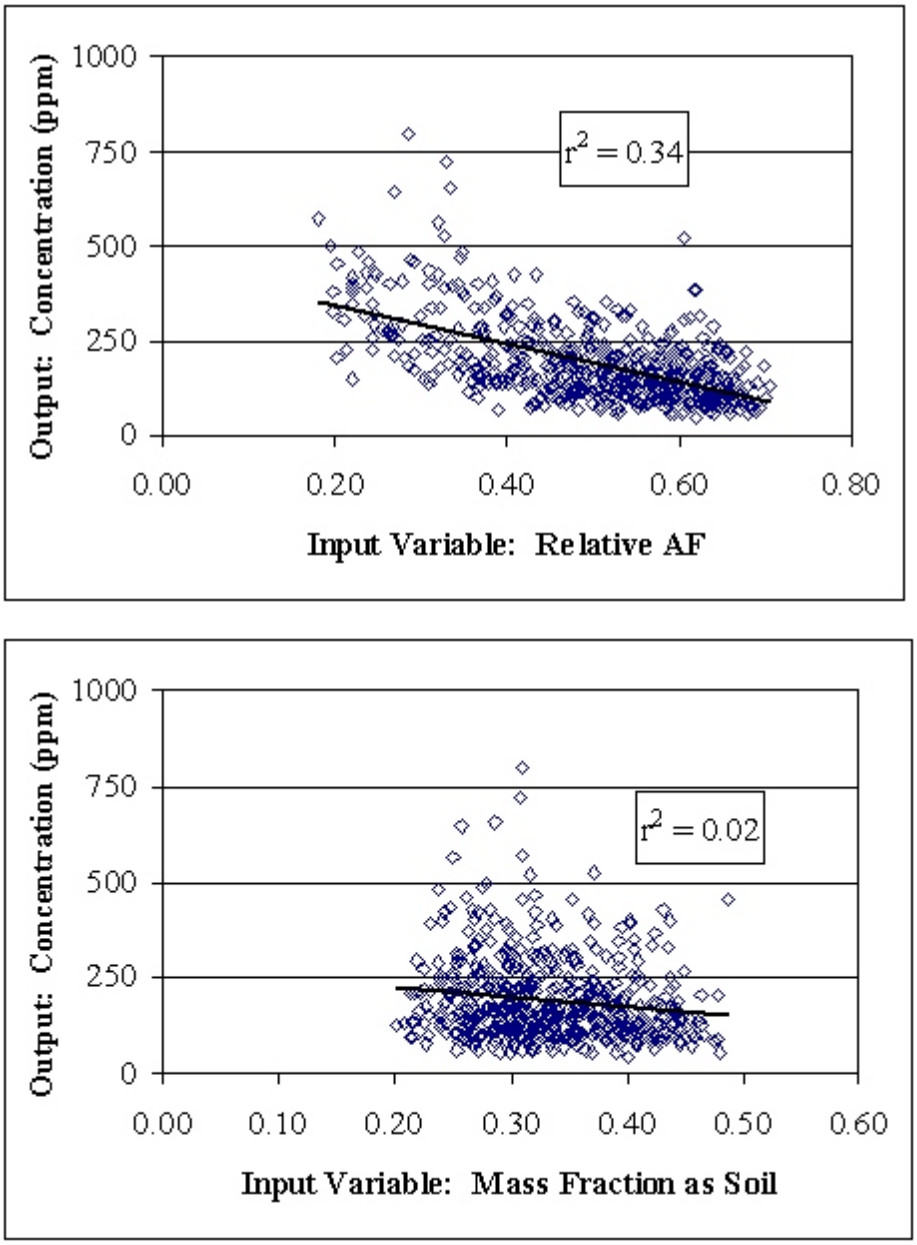


Figure A-2. Scatterplots of simulated random values from a 1-D MCA of variability. The output from the model is a contaminant concentration in soil (C) that corresponds with a prescribed (fixed) level of risk for a hypothetical population (based on Stern, 1994). For each iteration of a 1-D MCA simulation, random values were simultaneously selected for all model variables and the corresponding concentration (C) was calculated. Inputs were simulated as independent random variables. Scatterplots of 500 consecutive random values and estimates of C are shown for two input variables: relative absorption fraction, RAF (top graph); and mass fraction of dust as soil, F (bottom graph). There is a moderate, indirect relationship between C and RAF (r^2=0.34), compared with the weak relationship between C and F (r^2=0.02), suggesting that the model output (C) is more sensitive to variability in RAF than F.

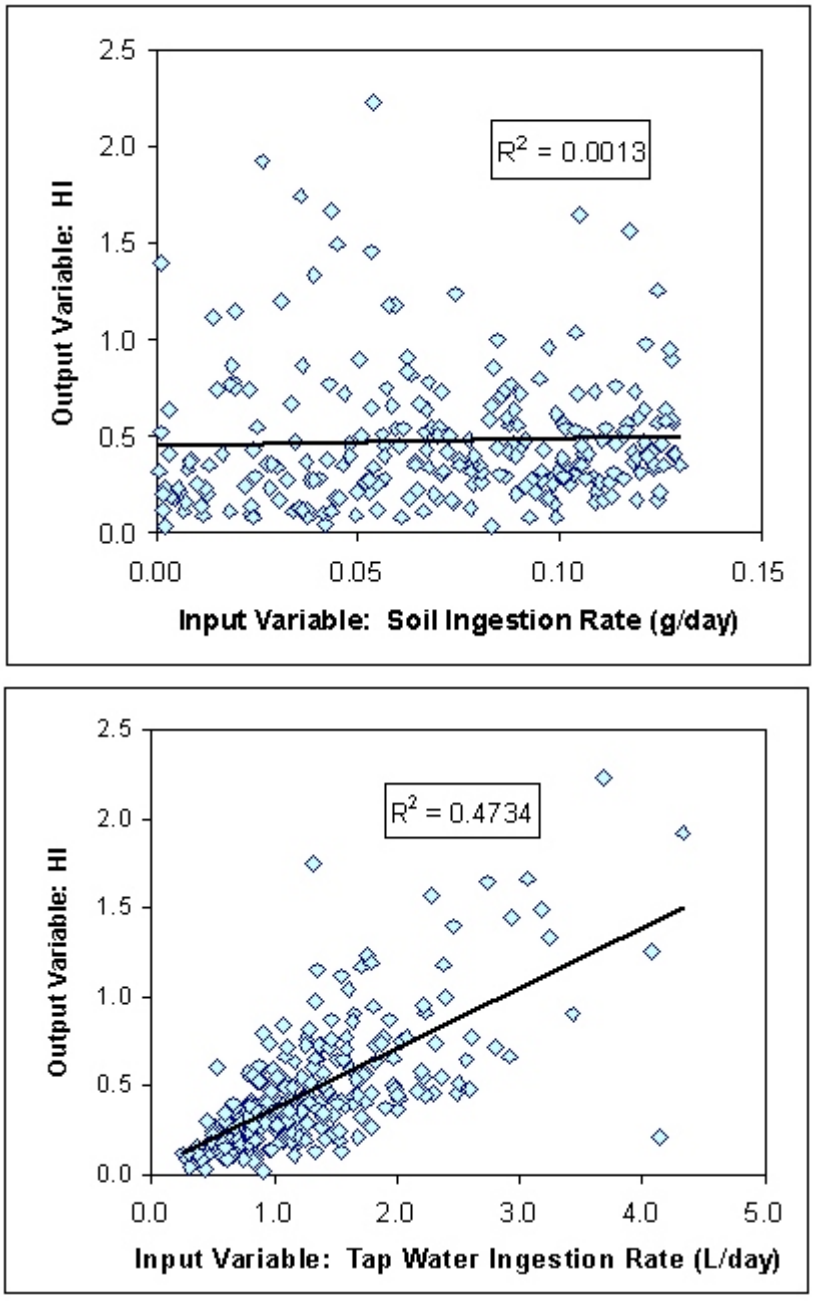


Figure A-3. Scatterplots of simulated random values from a 1-D MCA of variability for example in Section A.2.0. The output from the model is HI. For each iteration of a 1-D MCA simulation, random values were simultaneously selected for all model variables and the corresponding HI was calculated. Inputs were simulated as independent random variables. Scatterplots of 250 consecutive random values and estimates of HI are shown for two input variables: soil ingestion rate, I_s (top graph); and tap water ingestion rate, I_w (bottom graph). There is a negligible relationship between HI and I_s ($r^2 < 0.01$), compared with the strong relationship between HI and I_w ($r^2=0.47$), suggesting that the model output (HI) is more sensitive to variability in I_w than I_s. Best-fit lines were generated with the Simple Linear Regression in Microsoft Excel's trendline option for scatterplots; r^2 values represent the coefficient of determination (see Section A.3).

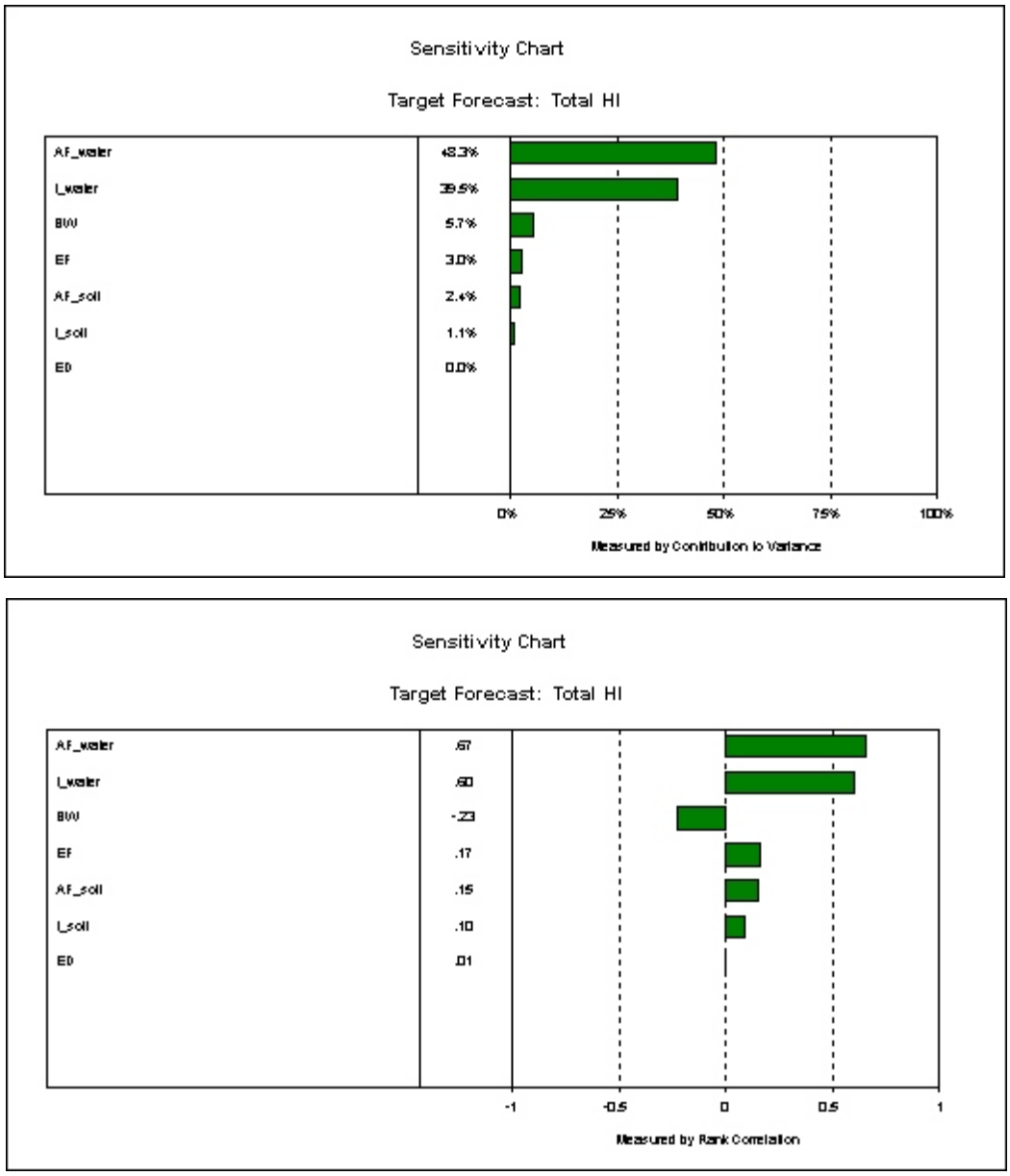


Figure A-4. Top panel - bar graph showing the r^2 values (square of Spearman rank correlation coefficient), a metric for the dependence of HI on exposure factors based on 1-D MCA for variability. Bottom panel - bar graph, sometimes referred to as "tornado plot", showing rank correlation coefficient. This graph is effective for showing both the relative magnitude and direction of influence (positive or negative) for each variable. Abbreviations for input variables are given in Table A-4. In this example, the variable with the greatest effect on HI is the absorption fraction in water (AF_w), followed by the water ingestion rate (I_w). Concentration does not influence variability because, in this example, long-term average concentration is characterized by a point estimate (i.e., 95% UCL), rather than a probability distribution. Exposure duration does not influence variability because variability in ED is expressed in both the numerator (ED) and denominator (AT=ED x 365 for noncarcinogenic effects), and cancels out. Output was generated with *Crystal Ball*®, which calculates the contribution to variance by squaring the rank correlation coefficient and normalizing to 100%.

In this example, seven exposure variables are used to characterize variability in HI. The remaining variables in the risk equation (i.e., concentration terms, and RfD) are characterized by point estimates. Because point estimates do not vary in a Monte Carlo simulation, they do not contribute to the variance in the output. This result does not mean that concentration is an unimportant variable in the risk assessment. Concentration may still contribute greatly to the uncertainty in the risk estimate. A sensitivity analysis of parameter uncertainty in a risk equation can be explored using iterative simulations, such as with 2-D MCA.

Results of the Pearson correlation and Spearman rank correlation give similar rankings of the input variables, with absorption fraction of water (AF_w) and tap water ingestion rate (I_w) being the two dominant exposure variables. Pearson correlations suggest that I_w is the most sensitive variable (r =0.644), whereas the highest Spearman rank correlation is for AF_w (r = 0.603). This may reflect the fact that I_w is characterized by an untruncated lognormal distribution, whereas AF_w is bounded between 0 and 1.0. The effect on the correlations of the occasional high-end value for I_w generated from random sampling of the lognormal distribution will tend to be expressed by Pearson correlations, but muted by the Spearman rank correlations.

A comparison of the Tier 1 and Tier 2 results is given below:

- **Tier 1, Sensitivity Ratios:**

  - Local SR (Δ = 5%) rankings: EF > BW > I_w = AF_w > I_s = AF_s > ED

  - Range SR (Δ = 50%) rankings: EF > I_w = AF_w > BW > I_s = AF_s > ED

- **Tier 1, Sensitivity Scores:**

  - Score based on local SR (Δ = 5%): I_w > AF_w > BW > EF > AF_s > IR_s > ED

  - Score based on range SR (Δ = 50%): I_w > AF_w > EF > BW > AF_s > IR_s > ED

- **Tier 2, Correlation Coefficients:**

  - Pearson: I_w > AF_w > BW > EF > AF_s > IR_s > ED

  - Spearman Rank: AF_w > I_w > EF > BW > AF_s > IR_s > ED

The Tier 1 sensitivity scores and Tier 2 correlation coefficients yield similar results, suggesting that, if sufficient information is available to estimate the coefficient of variation in the input variables, a Tier 1 analysis can help to focus efforts on the variables that contribute most to the variance in risk. By contrast, the Tier 1 sensitivity ratio approach suggested that EF was the most influential variable, when in fact it contributes less than 5% to the variance in the HI. These results suggest that Tier 1 sensitivity ratios are best applied to identify dominant exposure pathways, rather than dominant exposure variables in the risk equation.

A.2.2.3 FOCUSING ON THE RME RANGE OF THE RISK DISTRIBUTION

Monte Carlo methods can also be used to determine the sensitivity over a subset of the output distribution, such as the RME range (i.e., 90th to 99.9th percentiles). For some exposure models, the relative contribution of exposure variables may be different for the high-end exposed individuals than for the entire range of exposures. The general strategy for exploring sensitivity over subsets of risk estimates is to first sort the distribution of simulated output values in ascending (or descending) order, and then apply a sensitivity analysis to the subset of interest (e.g., > 90th percentile). For the hypothetical example presented in this appendix, there was no difference in the relative rankings of inputs in the RME range.

A.2.2.4 INSPECTION

With Monte Carlo analysis, the probability distributions assumed for the various input variables are used to generate a sample of a large number of points. Statistical methods are applied to this sample to evaluate the influence of the inputs on the model output. A number of different "indices" of sensitivity can be derived from the simulated sample to quantify the influence of the inputs and identify the key contributors. Most of these are based on an assumption that the model output *Y* varies in a monotonic, linear fashion with respect to various input variables (X_1, X_2, etc.). For example, an estimate of average daily intake (mg/kg-day) from multiple exposure pathways is linear with respect to the intake from each pathway. Since most risk models are linear with respect to the input variables, the output distribution (particularly its upper percentiles) tends to be dictated by the input variables with the largest coefficient of variation (CV), or the ratio of the standard deviation to the mean. For example, Equation A-9 represents a simple expression for intake rate as a function of random variables X_1 and X_2 :

$$Y = X_1 + X_2 \qquad \text{Equation A-9}$$

where X_1 and X_2 may represent dietary intake associated with prey species 1 and 2, respectively. If the same probability distribution was used to characterize X_1 and X_2, such as a lognormal distribution with an arithmetic mean of 100 and standard deviation of 50 (i.e., CV=50/100=0.5), each variable would contribute equally to variance in Y. If, however, X_2 was characterized by a lognormal distribution with an arithmetic mean of 100 and standard deviation of 200 (i.e., CV=200/100=2.0), we would expect Y to be more sensitive to X_2. That is, X_2 would be a greater contributor to variance in Y.

While the coefficient of variation may be a useful screening tool to develop a sense of the relative contributions of the different input variables, a common exception is the case when X_1 and X_2 have different scales. For example, Equation A-10 is an extension of Equation A-9:

$$Y = a_1 X_1 + a_2 X_2 \qquad \text{Equation A-10}$$

where a_1 and a_2 are constants that may represent the algebraic combination of point estimates for other exposure variables. If the means of X_1 and X_2 are equal, but $a_1 >> a_2$, then X_1 would tend to be the dominant contributor to variance, regardless of the CV for X_2. This concept was demonstrated by the sensitivity score calculations given in Table A-8. Water ingestion rate (I_w) and soil ingestion rate (I_s) had the same CV (0.58), but I_w was the dominant variable because tap water ingestion contributed approximately 90% to the HI.

$$Y = a_1 X_1 + a_2 X_2^{\theta}$$ Equation A-11

The most influential random variables generally have the highest degrees of skewness or are related to the output according to a power function (Cullen and Frey, 1999). For example, Equation A-11 presents an extension of Equation A-10 in which there is a power relationship between X_2 and Y. In this example, assume Y represents the total dietary intake rate of cadmium for muskrats, X_1 and X_2 represent the dietary intake rate associated with prey species 1 and 2, respectively, a_1 and a_2 represent additional point estimates in the equation, and θ is the power exponent. In general, for $\theta > 1$, the total dietary intake rate (Y) will be more sensitive to the intake rate associated with species 2 (X_2) than species 1. Assume (hypothetically) that the power relationship stems from the fact that there is a direct relationship between availability of prey species X_2 and chemical body burdens of prey species X_2 because individuals that are more accessible to the muskrat also happen to frequent areas of the site with higher concentrations.

A.3.0 Advanced Concepts in Sensitivity Analysis

This section provides additional information on the underlying principles of sensitivity analysis, although it is not a comprehensive summary and is not intended to substitute for the numerous statistical texts and journal articles on sensitivity analysis. Section A.3.1 begins with a general framework for relating model output to model input. Section A.3.2 explains the sensitivity ratio approach and highlights some of its limitations. Section A.3.3 reviews some of the metrics reported by the commercial software that report results of sensitivity analysis following Monte Carlo simulations (e.g., *Crystal Ball®*, *@Risk*). While statistical software for MCA provides convenient metrics for quantifying and ranking these sources, it is strongly recommended that risk assessors and risk managers develop an understanding of the underlying principles associated with these metrics.

A.3.1 Relating the Change in Risk to the Change in Input Variable X

For purposes of discussion, let *Y* denote a model output (e.g., risk) and suppose that it depends on the input variable *X*. In general, a risk assessment model may use any number of inputs; however, for purposes of illustrating concepts, it is convenient to restrict this discussion to one variable. The model relates the output *Y* to values of *X* (i.e., x_0, x_1, ♦♦♦, x_n) based on the function expressed as *Y=F(x)*. The sensitivity of *Y* to *X* can be interpreted as the slope of the tangent to the response surface *F(X)* at any point x_i. This two-dimensional surface can be a simple straight line, or it may be very complex with changing slopes as shown in Figure A-5a. The sensitivity, therefore, may depend on both the value of *X* and the amount of the change Δx about that point. This concept can be extended to two input variables, X_1 and X_2, where the response is characterized by a three-dimensional surface. The shape may be a simple plane (Figure A-5c) or it may be very complex with many "hills" and "valleys" depending on the defining function $F(X_1, X_2)$. In a typical risk assessment with ten or more variables, the surface can be very complex, but the shape is likely to be dominated by a small subset of the input variables.

A sensitivity analysis based on a relatively small deviation about the point may be referred to as a local sensitivity analysis, while a large deviation may be referred to as range sensitivity analysis. In either case, the objective is to evaluate the sensitivity at some nominal point (X_1*, X_2*) such as the point defined by the mean or median of X_1 and X_2. At any point, the sensitivity of the model output, $Y^* = F(X_1^*, X_2^*)$, to one of the inputs (X_1 *or* X_2), is represented by the rate of change in *Y* per unit change in *X*.

This is the slope of the surface at that nominal point in the direction of X and is expressed as $\partial Y/\partial X_i$, the *partial derivative* of Y with respect to X.

$$Partial\ Derivative = \frac{\partial Y}{\partial X} \approx \frac{\Delta Y}{\Delta X}$$

If the function $F(X_1, X_2)$ is known explicitly, it may be possible to determine the partial derivatives analytically. This is not a requirement, however, because an estimate can be obtained by incrementing X_i by a small amount, ΔX_i, while keeping the other inputs fixed and reevaluating the model output Y. The resulting change in Y divided by ΔX_i will approximate $\partial Y/\partial X_i$ at the nominal point. In practice, analytical solutions can be approximated using Monte Carlo techniques. This information is presented to highlight the fundamental concepts of sensitivity analysis. The partial derivative, *per se*, would typically not be one of the methods of sensitivity analysis used in a PRA. However, all of the approaches that are presented in this appendix are variations on this concept.

One drawback to using the partial derivative to quantify the influence of X_i is that the partial derivative is influenced by the units of measurement of X_i. For example, if the measurement scale for X_i is changed from grams to milligrams, the partial derivative $\partial Y/\partial X_i$ will change by a factor of 1,000. Therefore, it is necessary to **normalize the partial derivative** to remove the effects of units (see Section A.3.2).

If the relationship between Y and all of the inputs is linear, then the response surface is a flat plane and each of the partial derivatives at each point, (X_i, Y), will remain constant regardless of where the point is in the surface (Figure A-5b). In this case, it is a simple matter to determine the relative influence that the various inputs have on the model output. When the relationship is nonlinear, however, the situation is more complex because the influence of a particular input may vary depending on the value of that input.

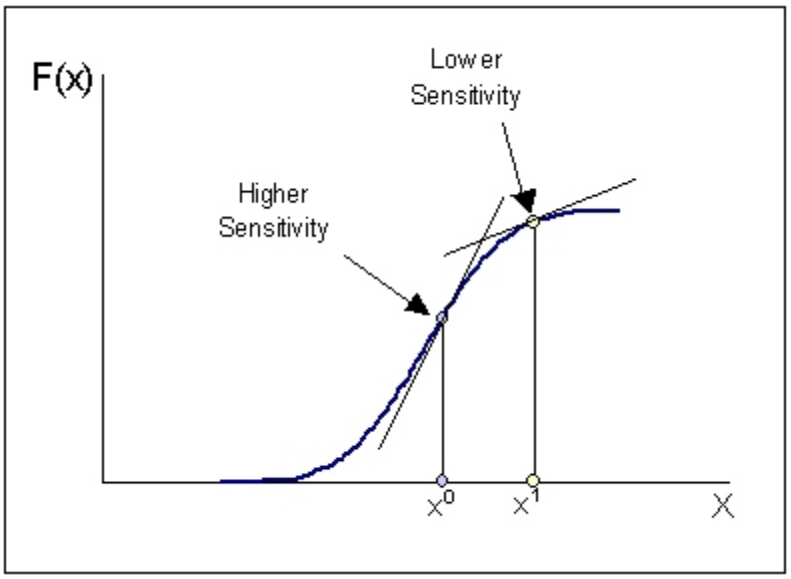


Figure A-5a. Hypothetical 2-D response surface for Y given one input variable: $Y=F(X)$. The sensitivity of Y with respect to X is calculated as the slope at a specific point on the surface (x^0, x^1), or the partial derivative, $\partial Y/\partial X_i$.

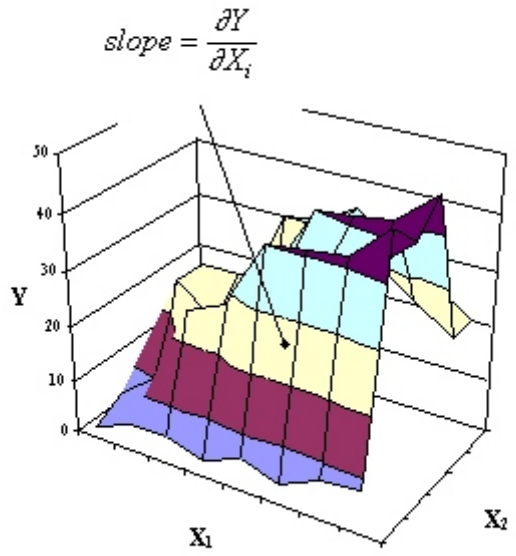


Figure A-5b. Hypothetical 3-D response surface for *Y* given two input variables: Y = $f(X_1, X_2)$. The sensitivity of *Y* with respect to X_i is calculated as the slope at a specific point on the surface, or the partial derivative, $\partial Y/\partial X_i$.

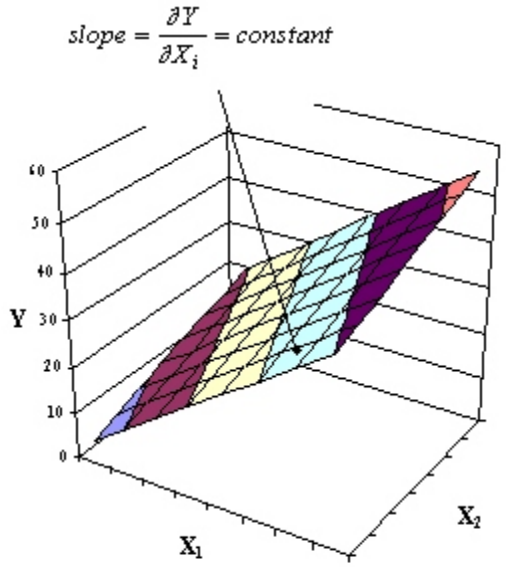


Figure A-5c. Hypothetical 3-D response surface when *Y* is a linear function of two input variables: Y=$f(X_1, X_2)$. The slope (i.e., the partial derivative, $\partial Y/\partial X_i$) is constant for any point (X_i, *Y*) on the surface in the direction of X_i. In this case, $\partial Y/\partial X_1$=5 while $\partial Y/\partial X_2$=2.

A.3.2 NORMALIZED PARTIAL DERIVATIVE

Classical sensitivity analysis methods use estimates of the partial derivatives of the model output with respect to each variable. For the purpose of evaluating the relative influence of the various input variables on the model output at a single point, the **normalized partial derivative** provides a useful index.

If the input variables are all discrete and take on a small number of values, then it is possible to evaluate the influence of the various input variables at each of the points defined by considering all possible combinations of the inputs. Then the influence can be evaluated for each input by computing normalized partial derivatives at each point. This approach is limited to situations where the number of inputs as well as the number of possible values for each input is relatively small; otherwise, the number of combinations to be evaluated will be unmanageable. Furthermore, when evaluating the influence at different points on the input-output surface simultaneously, it is important to take into account the probability associated with each of those points. For example, the fact that a particular input has a large influence on the model output at a particular point would be discounted if the probability associated with that particular point is very low.

A similar approach may be used to analyze inputs that are continuous variables if a few points representing the range of values are selected. For example, low, medium (or nominal), and high values may be selected for each of the continuous input variables and then the relative influence of each of the input variables can be computed as in the case of discrete inputs. One limitation of this approach, however, is that the continuous nature of the inputs makes it impossible to calculate an exact probability for each of the points. Generally, in a PRA, many if not all of the inputs will be random variables described by probability distributions and it will be necessary to quantify the influence of each input, X_i, over the entire range of X_i.

An estimate of the partial derivative can be obtained by incrementing X_i by a small amount, say ΔX_i while keeping the other inputs fixed and reevaluating the model output Y. The resulting change in Y divided by ΔX_i will approximate $\partial Y/\partial X_i$ at the nominal point.

$$\text{Partial Derivative} = \frac{\partial Y}{\partial X} \approx \frac{\Delta Y}{\Delta X}$$

As previously noted, one complication to using the partial derivative to quantify the influence of X_i is that the partial derivative is influenced by the units of measurement of X_i. One way this is accomplished is to divide the partial derivative by the ratio of the nominal point estimates, Y^* / X_i^* (or equivalently multiply by X_i^* / Y^*). An approximation of the normalized partial derivative is given by Equation A-12.

$$\text{Normalized Partial Derivative} \approx \frac{\Delta Y}{\Delta X} \times \frac{X_1}{Y_1} = \frac{\left(\frac{Y_2 - Y_1}{Y_1}\right)}{\left(\frac{X_2 - X_1}{X_1}\right)}$$ Equation A-12

This is the same as the equation for calculating sensitivity ratios (Section A.2.1.3), or elasticity (see Equation A-5). As with the SR approach, the normalized partial derived can be weighted by characteristics of the input variable (Section A.2.1.4). One approach is to divide by the ratio of standard deviations (σ_Y / σ_X), where σ_Y is the standard deviation of *Y* and σ_X is the standard deviation of *X*. This method requires that the standard deviations be known, or that a suitable estimate can be obtained.

As previously noted, if the relationship between *Y* and all of the inputs is nonlinear, the influence of a particular input may vary depending on the value of that input. One approach to this problem is to consider a range of values for the input and to examine the influence over that range. If the input is considered to be a random variable following some specified probability distribution, then it may be desirable to look at the influence that the random input has on the model output across the distribution of input values. This can be accomplished with a Monte Carlo approach. Another technique that addresses nonlinearities is to calculate contributions to variance using input variables that are transformed (e.g., lognormal or power transformation).

A.3.3 Regression Analysis: R^2, Pearson R, and Partial Correlation Coefficients

In order to understand R^2, it is necessary to first understand simple and multiple linear regression. In regression analysis, we are interested in obtaining an equation that relates a dependent variable (*Y*) to one or more independent variables (*X*):

$$Y = \beta_0 + \beta_1 X + \varepsilon$$ Equation A-13

where β_0 and β_1 are regression coefficients, and ε is called a random error. Equation A-13 is the general equation for a simple linear regression, because there is only one Y and one X variable, and their relationship can be described by a line with intercept β_0 and slope β_1.

Note that *linear* regression refers to the linear relationship between parameters (β_0, β_1), not *X* and *Y*. Thus, the equation $Y = \beta_0 + \beta_1 X_1^2 + \varepsilon$ is considered linear. *Multiple* linear regression involves more than one *X* related to one *Y* $\left[Y = \beta_0 + \beta_1 X_1 + \beta_2 X \cdots\right]$, while *multivariate* regression involves more than one *Y* to more than one *X*.

Exhibit A-5

Simplifying Assumptions in Regression Analysis

- *Y* is a linear function of the unknown coefficients (β_i)
- Successive values of *Y* are uncorrelated
- Variance of *Y* is constant for all values of inputs (X_i)

The random error, ε, represents the difference between an observed *Y* value (calculated from the observed input variables), and a Y value

predicted by the regression line ($\hat{y}$). It is also called the *residual* (i.e., $\varepsilon=y-\hat{y}$). The random error takes into account all unpredictable and unknown factors that are not included in the model. Exhibit A-5 gives some of the simplifying assumptions that apply to regression analysis. Assumptions about ε are that the random error has mean = 0 and constant variance, and is uncorrelated among observations. One method of finding the best regression line is to minimize the residual sum of squares (i.e., least-squares method), also called the sum of squares due to error (SSE).

In terms of sensitivity analysis, we are interested in how much of the variation in *Y* can be explained by the variation in *X*, and how much is unexplained (due to random error). If a scatter plot of paired observations (*x*, *y*) shows that our regression line intersects all of the observations exactly, then all of the variation in *Y* is explained by *X*. Another way of stating this is that the difference between the mean output ($\bar{y}$) and an observed *y* (y_i), or ($y_i - \bar{y}$), is equal to the difference between the mean output and a predicted *y* or ($\hat{y} - \bar{y}$).

In general, the total deviation of y_i from $\bar{y}$ is equal to the sum of the deviation due to the regression line plus the deviation due to random error:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$
$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \qquad \text{Equation A-14}$$
$$SST = SSE + SSR$$

Thus, the total sum of squares (SST) equals the sum of squares due to error (SSE) plus the sum of squares due to regression (SSR).

A.3.3.1 Calculations of R^2 and Adjusted R^2

The R^2 term is a measure of how well the regression line explains the variation in *Y*, or:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$
$$R = \sqrt{\frac{\text{variation explained by regression}}{\text{total variation in Y}}} \qquad \text{Equation A-15}$$

where R^2 is called the *coefficient of multiple determination* and *R* is called the *multiple correlation coefficient*. If R^2=0.90 for a certain linear model, we could conclude that the input variables (X_1, X_2,...X_k) explain 90% of the variation in the output variable (*Y*). R^2 reduces to the *coefficient of determination* r^2 for simple linear regression when one independent variable (*X*) is in the regression model. The *sample correlation coefficient*, *r*, is a measure of the association between *X* and Y, and calculated by Equation A-16. It is also referred to as the Pearson product moment correlation coefficient.

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2\right]^{0.5}}$$ Equation A-16

In addition, r is an estimate of the unknown population parameter, ρ, defined by Equation A-17:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$ Equation A-17

where σ_X and σ_Y denote the population standard deviations of the random variables X and Y, and where σ_{XY} is called the covariance between X and Y. The covariance σ_{XY} is a population parameter describing the average amount that two variables "covary". Thus, another way of thinking about a correlation coefficient (R) is that it reflects the ratio of the covariance between two variables divided by the product of their respective standard deviations; and the value always lies between -1 and +1. *@Risk* and *Crystal Ball®* provide both the R^2 for the entire model, as well as the correlation coefficients for each input variable (or regressor). The higher the value of R_i for X_i, the more sensitive the output variable is to that input variable.

Although the calculations are the same, there is a subtle conceptual difference between the coefficient of determination (r^2) from regression, and the square of the correlation coefficient. When evaluating two variables *(X, Y)*, the key is whether X is interpreted as a "fixed" quantity (i.e., an explanatory variable), or a random variable just like Y. In regression analysis, r^2 measures how well the regression line explains the variation in Y given a particular value for X (Equation A-15). Correlation requires that X be considered a random variable, typically having a bivariate normal distribution with Y (see Appendix B).

One artifact of regression analysis is that R^2 increases as you add more and more input variables to your model; however, the increased fit of the model due to one or more of the input variables may be insignificant. Sometimes an adjusted R^2 is calculated to take into account the number of input variables (called regressors) in the model (k) as well as the number of observations in the data set (n):

$$R^2_{adj} = \frac{(n-1)R^2 - 1}{n - k - 1}$$ Equation A-18

While R^2 gives the proportion of the total *variation* of Y that is explained, R^2_{adj} (Equation A-18) takes into account the degrees of freedom (*df*), and gives the proportion of the total *variance* of Y that is explained (variance = variation /df); or stated simply, R^2_{adj} is the R^2 corrected for *df*, where *df* is described by [1 - k/(n-1)].

- If the relationship between an input variable and an output variable is strong, but nonlinear, the R^2 statistic will be misleadingly low.

- If the means of the sampling data are used rather than the individual observations for each variable, R^2 will be misleadingly high. This is because taking the mean of a sample reduces the fraction of the *total* variation due to *random* variation (see discussion of random error above). This is an important consideration when trying to interpret the results of regression analyses that incorporate data averaged over different spatial scales (e.g., regression of PbB on soil lead concentrations taken at the city block level may give an inflated R^2 value if the sampling data are averaged over a larger spatial scale, such as the census tract level).

A multiple regression analysis can also be performed to estimate the **regression coefficients** (see Appendix A.3.3). Each coefficient essentially represents an "average" value of the partial derivative across the entire distribution of the input. The regression coefficient, like the partial derivative, depends on the units of measurement so, as in the case of the partial derivative, it must be normalized. This can be accomplished by multiplying the regression coefficient by the ratio of estimated standard deviations s_y/s_x.

A convenient way to carry out a sensitivity analysis is to perform a stepwise regression analysis. Some statistical software packages (e.g., SAS, SPSS) offer a variety of different approaches for this; however, in general, they can be classified into two general categories: forward selection and backward elimination. In the forward selection, the inputs are added to the model one by one in the order of their contribution. In the backward elimination, all of the inputs are used in the model initially and then they are dropped one by one, eliminating the least important input at each step. A true stepwise procedure is a variation on the forward selection approach where an input can drop out again once it has been selected into the model if at some point other inputs enter the model that account for the same information.

A.3.3.2 RELATIVE PARTIAL SUM OF SQUARES (RPSS)

The **relative partial sum of squares (RPSS)** measures the sensitivity of the model output to each of the input variables by partitioning the variance in the output attributable to each variable using multiple regression techniques (Rose et al., 1991). The RPSS is presented as a percentage reflecting the proportion of influence a given variable has on risk. The results of RPSS are intuitive and generally easy to understand.

Briefly, the RPSS represents the percentage of the total sum of squares attributable to each of the variables. To calculate RPSS for variable V_i, the difference between the regression sum of squares (RSS) for the full model and the regression sum of squares for the model with V_i missing (RSS_{-i}) is divided by the total sum of squares (TSS) and expressed as a percentage:

$$RPSS_i = \frac{100\,(RSS - RSS_{-1})}{TSS} \qquad \text{Equation A-19}$$

This procedure can be thought of as analogous to least squares linear regression, but performed in the *n*-dimensional parameter space of the risk equation. Since this approach depends on the adequacy of the linear regression model between the output variable (e.g., risk) and all the variables, an additional diagnostic is to check how close R^2 is to 1.0. For equations with more than three parameters (such as those used in Superfund risk assessments), the computational overhead of this process is large and requires specific computer programs. The software program *Crystal Ball*® does not perform this calculation, but it can be determined with most standard statistical software packages that perform

multiple regression. *@Risk* performs a calculation similar to this called multivariate stepwise regression that yields correlation coefficients in lieu of percent contributions to output variance.

A.3.3.3 SPEARMAN'S RANK CORRELATION COEFFICIENT (RHO)

The validity of using indices such as regression coefficients, correlation coefficients, and partial correlation coefficients depends on the assumptions of the underlying linear model being met. If there is any doubt that a data set satisfies the model assumptions, a nonparametric measure of correlation based on the rank orders of the inputs and associated outputs can be used. The Spearman Rank correlation coefficient is a nonparametric statistic; it measures an association between variables that are either count data or data measured on an ordinal scale, as opposed to data measured on an interval or ratio scale. An example of an ordinal scale would be the ranking of sites based on their relative mean soil concentrations. For example, if there are four categories of soil contaminant concentrations, sites with the highest concentrations may receive a rank of 1 while sites with lowest concentrations may receive a rank of 4. Ordinal scales indicate relative positions in an ordered series, not "how much" of a difference exists between successive positions on a scale.

To calculate the Spearman rank correlation coefficient, assign a rank to each of the input variables (X_j) and output variables (Y_k). For each ranked pair (X_j, Y_k), calculate the difference, d, between the ranks. For example, if the first observation for variable X has a ranking of 5 (relative to all of the observations of X), and the corresponding value of Y has a ranking of 3 (relative to all of the observations of Y), the difference (d) is equal to 5–3=2. Spearman rho (r_s) is calculated as:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n^3 - n)} \qquad \text{Equation A-20}$$

Hence ($-1 \le r_s \le 1.0$), and r_s=-1 describes a perfect indirect or negative relationship between ranks in the sense that if an X element increases, the corresponding Y element decreases. Similarly, r_s=0 suggests that there is no relationship between X and Y.

The Pearson product moment correlation coefficient is equal to the Spearman rank correlation coefficient when interval/ratio values of the measured observations (X, Y) are replaced with their respective ranks.

REFERENCES FOR APPENDIX A

Cullen, A.C. and H.C. Frey. 1999. Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs. Plenum Press.

Hamby, D.M. 1994. A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models. *Environ. Monit. and Assess.* 32:135–154.

Helsel, D.R. and R.M. Hirsch. 1992. Statistical Methods in Water Resources. Elsevier Science B.V.

Iman, R.L. and J.C. Helton. 1988. An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Risk Anal.* 8:71–90.

Iman, R.L. and J.C. Helton. 1991. The Repeatability of Uncertainty and Sensitivity Analyses for Complex Probabilistic Risk Assessments. *Risk Anal.* 11:591–606.

Merz, J., M.J. Small, and P. Fischbeck. 1992. Measuring Decision Sensitivity: A Combined Monte Carlo-Logistic Regression Approach. *Medical Decision Making*, 12: 189–196.

Morgan, M.G. and M. Henrion.1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Palisade Corporation. 1994. *Risk Analysis and Simulation Add-In for Microsoft Excel or Lotus 1-2-3.* Windows Version Release 3.0 User's Guide, Palisade Corporation, Newfield, NY.

Rose, K.A., E.P. Smith, R.H. Gardner, A.L. Brenkert, and S.M. Bartell. 1991. Parameter Sensitivities, Monte Carlo Filtering, and Model Forecasting Under Uncertainty. *J. Forecast* 10:117–133.

Saltelli, A and J. Marivort. 1990. Non-Parametric Statistics in Sensitivity analysis for Model Output: A Comparison of Selected Techniques. *Reliab. Engin. Syst. Saf.* 28:299–253.

Shevenell, L. and F.O. Hoffman. 1993. Necessity of Uncertainty Analyses in Risk Assessment. *J Hazard Mater.* 35:369–385.

Stern, A.H. 1994. Derivation of a Target Level of Lead in Soil at Residential Sites Corresponding to a *de minimis* Contribution to Blood Lead Concentration. *Risk Anal.* 14:1049–1056.

U.S. EPA. 1997. *Guiding Principles for Monte Carlo Analysis.* Risk Assessment Forum and National Center for Environmental Assessment. EPA/630/R-97/001.

U.S. EPA. 1999. *TRIM, Total Risk Integrated Methodology, TRIM FATE Technical Support Document Volume I: Description of Module*. Office of Air Quality Planning and Standards. EPA/43/D-99/002A.